DOCUMENT RESUME

ED 269 464 TM 860 318

AUTHOR Holmes, Susan E.
TITLE Multi-Purpose Tests: A Solution to Test Proliferation.
PUB DATE 19 Apr 86
NOTE 50p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 17-19, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Basic Skills; *Content Analysis; Correlation; Elementary Secondary Education; Field Studies; Grade 6; Item Analysis; Latent Trait Theory; Mathematics Tests; Norm Referenced Tests; *Research Methodology; Scaling; *Testing Problems; *Testing Programs; *Test Use; Test Validity
IDENTIFIERS *California Assessment Program; *Stanford Achievement Tests; Test Redundancy

ABSTRACT

This research, conducted for California State Department of Education, is one component of a feasibility study focusing on ways to make the testing process more efficient by application of multi-purpose tests (MPT). MPT are designed, administered and scored to serve more than one purpose and to consolidate and unify testing programs as a partial solution to test proliferation. A content review identifies common skills measured by the California Assessment Program (CAP) and the Stanford Achievement Test. Empirical information was collected about matched CAP and Stanford item performance from a mathematics field test. A core set of multi-purpose items was identified. A validation study of over 1,000 grade 6 students examined the accuracy of the estimates of norm-referenced test information from the multi-purpose item set identified. The general strategy was to estimate Stanford total test and subtest scores in mathematics by applying item response theory techniques to the identified multi-purpose item set. Estimated scores were then compared with total test and subtest scores actually earned by students. The correlations between pairs of estimated and actual scaled scores were large, and scaled root mean square differences were small. Results indicate: (1) some testing programs evaluate common skills, supporting the concept of a consolidated testing system; (2) there is a need for statewide content domains; and (3) using multi-purpose testing methodology is feasible. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269464

Multi-Purpose Tests:
A Solution to Test Proliferation

Susan E. Holmes, Ph.D.
The Psychological Corporation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. E. Holmes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A paper presented at the annual meeting of the
National Council on Measurement in Education in San Francisco
April 19, 1986

TM 860 318

Abstract

The nation's investment in school achievement testing has continued to grow in the 1980's. The burden of this increase is most directly felt by students and classroom teachers who must relinquish precious instructional time for testing. The research reported here is one component of a feasibility study, conducted for the California State Department of Education. It focuses on ways to make the testing process more efficient by designing tests that serve more than one purpose. The study reports on the identification of common skills assessed in two testing programs based on a content review, and the subsequent development and validation of a prototype multi-purpose test in mathematics which serves both state and local information needs. The study highlights the need for well-defined content domains that include all important outcomes of the educational programs being assessed.

Multi-Purpose Tests:
A Solution to Test Proliferation

Introduction

Testing issues continue to confront the educational community in the 1980's as the nation's investment in school achievement testing continues to grow. Many factors have contributed to the proliferation of testing. Included among these are the public's demand for accountability, mandates for proficiency and minimum competency testing, judicial decisions defining increased responsibilities for public schools, and evaluation requirements for government-funded educational programs. The burden of this increase in testing is most directly experienced by students and classroom teachers. Both groups must relinquish large amounts of precious instructional time for testing.

The list of testing programs in which students and teachers participate each year can be a long one. There is usually a state-level testing program designed to collect information for monitoring student progress and for program evaluation at the state level. There are also district-level testing programs, frequently using norm-referenced standardized achievement tests, designed to provide information about the performance of individual students. This information is used to make national comparisons of student performance for reporting to school boards and parents and to satisfy federally mandated program evaluation requirements. The results of these testing programs are also used for diagnostic assessment of individual students. And finally, there may be still other district-level testing programs designed to evaluate student performance in terms of locally specified skills and objectives.

Opponents of testing have questioned the value of its contribution to the learning process for a number of years. They have now added to their list of objections the fact that many of the testing programs in which students and teachers participate seem to measure

the same things. While each program serves a different purpose and therefore has a different focus, overall, they appear to have a considerable amount of redundancy in terms of the content assessed. The issue seems to be one of emphasis rather than the total absence or presence of specific content (Jolly & Gramenz, 1984; Madaus, Kellaghan, Rakow, & King, 1979; Popham, 1978). The challenge then is to devise new ways to make tests, and the testing process, more efficient.

The research reported here focuses on this challenge. It describes an application of a concept referred to as "multi-purpose" tests. Multi-purpose tests are tests designed, administered, and scored in ways that satisfy more than one purpose or need. This concept is presented here as a potential solution to the problem of test proliferation. It is explored as a way to consolidate and unify state- or district-level testing programs. While certainly not a new concept, multi-purpose tests have the potential to reduce testing time, as well as increase the quality of the resulting test information. Either of these features makes the research described here worth considering at this time.

The Feasibility Study

The research reported here is part of a larger study conducted for the California State Department of Education. The current testing system in California has two major components: a state-level testing program, the California Assessment Program (CAP) and various district-level testing programs. The first component assesses a broad range of educational objectives in the state's core curriculum and provides a wide array of scores which characterize student achievement at the school, district, and state levels in a variety of subject areas. The second component consists of a variety of district-level testing programs. It includes, for the most part, standardized norm-referenced achievement testing programs. These programs provide districts with the information they need to satisfy program evaluation requirements, to make national comparisons of

student performance, and to report to parents and local school boards.

One of the major goals of California's educational reform movement is to consolidate and unify the various testing programs currently administered in public schools and to increase the usefulness of the testing information provided to students, teachers, parents, and the public. To accomplish this end, the State Department entered into a cooperative relationship with several test publishers to development prototypic models of such a system, referred to as the Comprehensive Assessment System (CAS), and to test the feasibility of these models. For a detailed description of the prototypic model proposed by The Psychological Corporation and other components of this study, see Volumes 1 and 2 of the Phase 1 Pilot Testing: Final Report (The Psychological Corporation, 1985).

The research reported here was conducted in four phases. The four phases are:

Phase 1: Content Review
Phase 2: Field Test
Phase 3: Identification of a Multi-Purpose Item Set
Phase 4: Validation Study.

Each phase will be summarized in the pages that follow.

Phase 1: Content Review

One of the basic assumptions of multi-purpose testing as defined here is that the many testing programs in California actually assess a large number of the same skills or objectives. In order to evaluate the tenability of this assumption, a review was conducted of the skills measured by the California Assessment Program and those measured by a standardized, norm-referenced achievement battery commonly used in district-level testing programs.

1. California Assessment Program, Grade 6 Mathematics. The curriculum commonly taught in public schools throughout California is enumerated in statewide curriculum frameworks. These have been translated into test content specifications and test items by committees of content specialists convened expressly for this purpose. This work was guided by the statewide curriculum frameworks and a content analysis of commonly used state-adopted textbook series. The resulting list of essential skills in reading, mathematics, and written language is assessed statewide each year by the California Assessment Program (CAP).

The list of essential skills in mathematics at Grade 6 contains 33 skills, referred to as "skill areas" or "skill elements" (The California State Department, 1980). A skill area is an "indivisible curricular element." Skill areas are narrowly defined so that any changes in curricular emphasis will have the same effect on items measuring them. Each skill area is measured by between 10 and 25 test items, making the total number of items in the mathematics domain 480. The 33 mathematics skill areas at Grade 6 and the number of items measuring each is presented in Table 1.

2. The Stanford Achievement Test. The norm-referenced standardized achievement test series used in the study was the Stanford Achievement Test, Seventh Edition (Gardner, Rudman, Karlsen, & Merwin, 1982). The mathematics skills assessed by the Stanford at Grade 6 were therefore also reviewed during the first phase of the study.

The Stanford consists of ten tests or levels, spanning Kindergarten through Grade 12. The purpose of the Stanford is to assess the performance of students in the basic skills areas of reading, mathematics, language arts, science, and social science. A common scaled score system is used to provide continuous assessment across all grade levels. The ten tests, the grade ranges for which each is appropriate, and their administration times are given in Table 2.

Table 1

The 33 Mathematics Skill Areas Measured by the California Assessment Program at Grade 6 and the Number of Items Used to Assess Each

Skill Area	Description	Number of Items
1. I.A.1.	Counting and Numeration: Skills	15
2. I.A.2.	Place Value: Skills	10
3. I.B.	Counting, Numeration and Place Value: Applications	15
4. II.A.1.	Ordering and Properties: Skills	15
5. II.A.2.	Classification of Numbers: Skills	20
6. II.B.	Nature of Numbers and Properties: Applications	15
7. III.A.1.	Addition/subtraction of Whole Numbers: Skills	15
8. III.A.2.	Multiplication of Whole Numbers: Skills	14
9. III.A.3.	Division of Whole Numbers: Skills	15
10. III.A.4.	Addition/subtraction of Decimals: Skills	14
11. III.A.5.	Multiplication/division of Decimals	12
12. III.A.6.	Operations on Fractions: Skills	16
13. III.A.7.	Percents and Equivalent Fractions and Decimals	12
14. III.B.1.	One-step Involving Whole Numbers: Applications	12
15. III.B.2.	One-step Involving Rational Numbers: Applications	20
16. III.B.3.	Two (or more) Steps: Applications	15
17. IV.A.1.	Expressions and Equations: Skills	15
18. IV.A.2.	Graphs and Function Tables: Skills	12
19. IV.B.	Expressions, Equations, Coordinate Graphs: Applications	15
20. V.A.1.	Shapes and Terminology: Skills	12
21. V.A.2.	Relationships: Skills	12
22. V.B.	Geometry: Applications	16
23. VI.A.1.	Metric Units: Skills	20
24. VI.A.2.	U.S. Customary Units: Skills	10
25. VI.A.3.	Perimeter, Area, and Volume: Skills	12
26. VI.B.	Measurement: Applications	16
27. VII.A.	Probability	2
28. VII.B.	Statistics	11
29. VIII.A.	Tables and Graphs	15
30. VIII.B.	Integrated Applications	15
31. IX.A.	Formulation: Problem Solving	15
32. IX.B.	Analysis and Strategy: Problem Solving	25
33. IX.C.	Interpretation: Problem Solving	12
	Total:	480

Table 2

The Ten Levels of the Stanford Achievement Test, Their Grade Ranges and Administration Times

Level	Grade Range	Administration Time
SESAT 1	K.0 - K.9	2 hr. 10 min.
SESAT 2	K.5 - 1.9	2 hr. 25 min.
Primary 1	1.5 - 2.9	3 hr. 35 min.
Primary 2	2.5 - 3.9	3 hr. 45 min.
Primary 3	3.5 - 4.9	4 hr. 55 min.
Intermediate 1	4.5 - 5.9	5 hr. 15 min.
Intermediate	5.5 - 7.9	5 hr. 15 min.
Advanced	7.0 - 9.9	4 hr. 40 min.
TASK 1	8.0 -12.9	3 hr. 5 min.
TASK 2	9.0 -13.0	3 hr. 5 min.

Note. Administration times are for the complete battery.

The Stanford series is designed to reflect what is currently being taught in schools throughout the country. In order to ensure that the test content is valid, test specifications, or blueprints, are based on a thorough analysis of curricular materials. The most widely used textbook series in various areas were studied, along with syllabi, state guidelines, and all pertinent research literature.

Test specifications for the Stanford were prepared in terms of instructional objectives. These test blueprints indicate the proportion of test content to be devoted to each topic to insure a balanced coverage of the curriculum. Blueprints were reviewed by curriculum specialists. Table 3 provides an outline of the mathematics content for the Intermediate 2 level as determined by the final review. Mathematics performance is generally reported in terms of three subtest scores and a total score. The subtest scores are Concepts of Number, Mathematics Computation, and Mathematics Applications. Cluster scores at the objective level are available. A more detailed description of the objectives measured in all levels of the series can be found in the "Index of Instructional Objectives" (Gardner, Rudman, Karlsen, & Merwin, 1983).

Table 3

Mathematics Objectives and Number of Items Measuring Each Objective in the Intermediate 2 Level of the Stanford Achievement Test

Subtest	Objective/Cluster	Number of Items
	Concepts of Number	34
	Whole Numbers and Place Value	12
	Recognize Names for Numbers and Read Numerals	(4)
	Interpret Numerals and Round Numbers	(5)
	Compare and Order Whole Numbers	(3)
	Rational Numbers	16
	Fractions	(8)
	Decimals	(6)
	Integers	(2)
	Operations and Properties	6
	Mathematics Computation	44
	Addition with Whole Numbers	4
	Subtraction with Whole Numbers	4
	Multiplication with Whole Numbers	6
	Division with Whole Numbers	10
	Division Facts	(3)
	Division by Numbers Less Than Ten	(4)
	Division by Two-Digit Divisors	(3)
	Computation with Fractions	8
	Computation with Decimals	6
	Estimation	3
	Proportions	3
	Mathematics Applications	40
	Problem Solving	22
	Solve Problems	(16)
	Solution Sentences	(3)
	Missing/Irrelevant Data	(3)
	Geometry/Measurement	12
	Geometry	(3)
	Measurement	(9)
	Graphs, Charts, and Tables	6
	Single Bar Graphs	(3)
	Tables	(3)
	Total Mathematics:	118

3. Item Mapping. In order to assess the degree of overlap between skills measured by the California Assessment Program (CAP) and those measured by the Stanford, items from the Stanford series were

reviewed. All potential matches for a given CAP skill area were identified. When the skill area was composed of subskills, the specific subskill measured was also identified.

The text of potentially matching Stanford items was then pasted on pages, along with selected item statistics, and organized into sets by CAP skill area. This procedure was also employed for all of the CAP items. The two sets of items purported to measure the same skill, CAP and Stanford, were then read and the correct response to each noted. Content was the prime consideration in finalizing the match, although item format and difficulty were also considered. Based on this review, a pool of 178 Stanford items was identified. The results of this item-by-item mapping are presented in Table 4.

Five of the skill areas assessed by the California Assessment Program are not measured in Stanford. The five skill areas are:

I.B.	Counting, Numeration, and Place Value: Applications
II.B.	Nature of Numbers and Properties: Applications
V.B.	Geometry: Applications
VI.A.2.	Measurement Skills: Customary Units
IX.C.	Problem Solving: Interpretation

Six more skill areas were measured by fewer than three Stanford items. The remaining 22 skill areas were each measured by between three and ten Stanford items. The average number of Stanford items measuring each skill area was roughly five items.

Most skill areas are composed of one or more specific subskills, each of which is measured by some items. Approximately 50 percent of these subskills are also measured in Stanford.

Table 4

The 33 Mathematics Skill Areas Measured by the California Assessment Program at Grade 6, the Number of Items Used to Assess Each, and the Number of Stanford Items Also Measuring that Skill

Skill Area		Description	Number of Items CAP	Number of Items Stanford
1.	I.A.1.	Counting and Numeration: Skills	15	7
2.	I.A.2.	Place Value: Skills	10	7
3.	I.B.	Counting, Numeration and Place Value: Applications	15	0
4.	II.A.1.	Ordering and Properties: Skills	15	9
5.	II.A.2.	Classification of Numbers: Skills	20	6
6.	II.B.	Nature of Numbers and Properties: Applications	15	0
7.	III.A.1.	Addition/subtraction of Whole Numbers: Skills	15	15
8.	III.A.2.	Multiplication of Whole Numbers: Skills	14	8
9.	III.A.3.	Division of Whole Numbers: Skills	15	12
10.	III.A.4.	Addition/subtraction of Decimals: Skills	14	6
11.	III.A.5.	Multiplication/division of Decimals	12	6*
12.	III.A.6.	Operations on Fractions: Skills	16	11
13.	III.A.7.	Percents and Equivalent Fractions/Decimals	12	8
14.	III.B.1.	One-step Involving Whole Numbers: Applications	12	5
15.	III.B.2.	One-step Involving Rational Numbers:	20	6
16.	III.B.3.	Two- (or more) Steps: Applications	15	10
17.	IV.A.1.	Expressions and Equations: Skills	15	9
18.	IV.A.2.	Graphs and Function Tables: Skills	12	2*
19.	IV.B.	Expressions, Equations, and Coordinate Graphs: Applications	15	6**
20.	V.A.1.	Shapes and Terminology: Skills	12	6*
21.	V.A.2.	Relationships: Skill	12	4
22.	V.B.	Geometry: Applications	16	0
23.	VI.A.1.	Metric Units: Skills	20	1
24.	VI.A.2.	U.S. Customary Units: Skills	10	0
25.	VI.A.3.	Perimeter, Area, and Volume: Skills	12	1
26.	VI.B.	Measurement: Applications	16	12
27.	VII.A.	Probability	12	1
28.	VII.B.	Statistics	11	1
29.	VIII.A.	Tables and Graphs	15	6
30.	VIII.B.	Integrated Applications	15	6
31.	IX.A.	Formulation: Problem Solving	15	6
32.	IX.B.	Analysis and Strategy: Problem Solving	25	1
33.	IX.C.	Interpretation: Problem Solving	12	0
		Total:	480	178

* Items from the Advanced Level

** Includes two items from the Primary 3 Level

Phase 2: Field Test

The second phase of the study consisted of a mathematics field test in which empirical information was collected about the performance of the matched CAP and Stanford item sets. Based on the data collected here, a set of multi-purpose items was identified in the third phase of the study.

Testing Materials

A variety of test materials were designed and produced by The Psychological Corporation during the second phase of the study. Included among these are the following:

1. Three field test booklets:
   Survey of Basic Mathematics Skills, Grade 6, Form A
   Survey of Basic Mathematics Skills, Grade 6, Form B
   Survey of Basic Mathematics Skills, Grade 6, Form C
2. A customized answer document;
3. A Scoring Service Identification Sheet;
4. A 1985 Test Examiner's Manual: Survey of Basic Mathematics Skills, Grade 6;
5. A 1985 Test Coordinator's Manual: Survey of Basic Mathematics Skills, Grade 6.

1. Field Test Booklets. Three field test forms were designed and produced for the field test. Each form contained two kinds of test items: CAP items and Stanford items. A total of 292 CAP items from 21 of the 33 skill areas measured by the California Assessment Program were included in the field test. One hundred thirty seven Stanford items were included.

Skill areas were selected for inclusion in the field test based on the availability of matching Stanford items and space limitations within field test forms. In most instances, all the CAP items measuring a skill area were included.

Design of Field Test Booklets. The 292 CAP items and the 137 Stanford items were configured into three field test forms, labeled Forms A, B, and C. For ease of administration, the items in each form were organized into four subtests, each of which could be administered in approximately one class period. Table 5 presents the four subtests in each form, the number of items in each subtest, and the testing time required to administer each subtest.

Table 5

The Four Subtests in Each Field Test Form, the Number of Items in Each Subtest, and the Time Required for Administration

Subtest	Number of Items	Administration Time
Concepts of Number	40	30 min
Mathematics Computation	45	40 min
Mathematics Applications: Part 1	38	35 min
Mathematics Applications: Part 2	30	30 min
		135 min

Skill areas were assigned to field test forms so that the difficulty of a particular subtest, as well as the overall difficulty of each form, would be as similar as possible. Initially, an average p-value was computed for each of the 21 skill areas. For the CAP items, these computations were based on values reported for the April 1984 administration of the California Assessment Program. For Stanford items, values from the spring 1982 standardization of the Stanford Achievement Test were used. An overall average was then used as an estimate of the difficulty of each subtest. The mean values obtained are reported in Table 6 for each skill area, subtest, and form of the field test.

Table 6

Skill Areas and Mean P-Values by Field Test Form

	Form A		Form B		Form C	
Subtest	Skill Area	Mean p	Skill Area	Mean p	Skill Area	Mean p
Concepts of Number	I.A.1.	.69	I.A.2.	.65	II.A.2.	.59
	III.A.7.	.64	II.A.1.	.66	IV.A.2.	.62
		.67		.66		.60
Computations	III.A.3.	.71	IIIA.2.	.78	III.A.1.	.77
	III.A.5.	.53	III.A.6.	.54	III.A.4.	.60
		.64		.65		.66
Applications: Part 1	IX.A.	.67	III.B.1.	.70	VIII.A.	.67
	V.A.2.	.63	VIII.B.	.65	V.A.1	.66
		.65		.68		.66
Applications: Part 2	VI.B.	.57	III.B.2.	.59	III.B.3.	.56
Overall Mean p-value		.64		.65		.64

Note. Mean p-values for skill areas are for the CAP and Stanford items combined.

Sequencing of Items Within Subtests. Once skill areas had been assigned to subtests within field test forms, the items measuring each skill area were sequenced within those subtests using one of two schemes. In the first scheme, employed for the Concepts of Number subtests only, all items measuring a given skill area were presented together, in sequential order from easiest to hardest, beginning with the easier skill area in that subtest. In this case, CAP and Stanford items were intermixed throughout the subtest.

In the second scheme, used for the remaining three subtests, all Stanford items appeared first in the subtest, arranged from easiest to hardest within skill area, beginning with the easier skill area. The Stanford items were then followed by the CAP items, again arranged from easiest to hardest within a skill area, beginning with the easier skill area. The second scheme for arranging items within subtests was necessary because of differences between the CAP and Stanford items in numbers of answer options. This problem is discussed more fully in the next section of the report.

Practice Questions. In order to give students an opportunity to practice their test taking skills, each subtest was preceded by a practice item. These items appeared in each field test form and were similar in format to the rest of the items in the subtest in which they appeared. Directions for answering practice items were read aloud by the test examiner, and students recorded their answers on the answer document. For the Mathematics Computations and both Mathematics Applications subtests, the practice item had five answer options, the last being "Not Here." At this time, the meaning of this answer choice was given and students had the opportunity to ask questions if they were unclear about its meaning.

2. Answer Document. A customized answer document was designed and produced for the study by The Psychological Corporation. Customization was necessary because all of the CAP items have four answer options, while the Stanford computations and applications items have five. In many instances, the fifth answer option is "Not Here." The difference in numbers of answer options can cause students to incorrectly record their answers if an answer document for five option items is used. This is particularly a concern in instances where the fourth option is the correct one. In this case, students will frequently record their answer in the bubble for the fifth.

3. Scoring Service Identification Sheet. A Scoring Service Identification Sheet was also part of the test materials for the field test. This pre-gridded document allowed information needed to scan and score the field test answer documents to be captured electronically.

4. Test Examiner's Manual. Specific instructions for test examiners were printed in the 1985 Test Examiner's Manual: Survey of Basic Mathematics Skills developed by The Psychological Corporation for the study. The test examiner's manual consisted of a description of the test materials sent to each classroom, procedures to follow prior to, during, and following the field test, and specific directions for the test administration. A copy of this manual was sent to each teacher participating in the field test.

5. Test Coordinator's Manual. Instructions for test coordinators were also prepared as part of the testing materials developed by The Psychological Corporation. These appeared in the 1985 Test Coordinator's Manual: Survey of Basic Mathematics Skills. They were designed to assist the test coordinators in carrying out their responsibilities related to the field test. The instructions included general information about the purpose of the field test, a description of the test materials for each classroom, a description of the role of the test coordinator, and a list of the activities to be completed prior to, during, and following the field test.

Field Test Sample

The field test was conducted in six California school districts. Over 60 individual schools participated in the testing of 4,171 students in 128 classrooms. The data from 3,470 of these students were analyzed. Approximately half the sample was male (n = 1770), and half the sample was female (n = 1700).

For the purposes of this study, students were included in the data analysis only if they had taken all four subtests and had attempted at least three of the first six items in each. The total number of valid cases was 1,138, 1,145 and 1,187, respectively, for Forms A, B, and C. Approximately half of each of these groups was male, and half was female.

Field Test Procedures

Assignment of Field Test Forms to Classrooms. All three field test forms were administered in each classroom. Field test forms were assigned to classrooms so that a more or less equal number of students from across the mathematics ability continuum took each form. The 1984 mean mathematics scale score earned by each school was used as an indicator of the mathematics ability level of the classrooms in that school. These values were made available by the California Assessment Program. The mean mathematics scale score for those students taking Forms A, B, and C, respectively, was 274.4, 269.3, and 271.2. The mean and standard deviation of this scaled score system is 250 and 50, respectively.

Test Administration. The field test was conducted during the third week in May in 1985. This time period was selected because it falls within the empirical norm dates for the Stanford and because it was very close to the spring testing conducted by the California Assessment Program.

Test materials, packaged by classroom, were shipped to the designated test coordinator in each district during the first week in May. In most instances, this individual was the person normally responsible for testing at the district level. The test coordinator then distributed the materials to participating classrooms. When testing was complete, materials were returned to the test coordinator, who then mailed the completed answer documents for that district to Cleveland for processing.

The test administration itself was conducted by classroom teachers. Teachers were instructed to administer the test on four different days during the testing period, preferably during the morning. If two subtests were administered on the same day, teachers were told that one subtest should be given in the morning and the other in the afternoon.

Teachers were instructed to administer the test to all students in their classroom who would normally be tested by the California Assessment Program. This program excluded students with mental or physical handicaps, as well as students who could not meaningfully interact with the English language. Make-up testing for students absent for one or more subtests was encouraged, although it was not considered mandatory.

Results

Raw Score Summary Statistics. The field test data were analyzed twice. The first analysis included both CAP and Stanford items. The second analysis focused on only CAP items. The latter analyses were performed so that comparisons could be made between the 1981-82 CAP statewide assessment results reported for these items and those obtained here. These analyses are not be reported here.

Tables 7 through 9 summarize the results of the field test in terms of raw scores. The total test, subtest, and skill area raw score means and standard deviations are reported there by field test form for combined CAP and Stanford item sets. The number of items comprising each score and the average percent correct are also presented.

The values reported in Tables 7 through 9 are based on combined CAP and Stanford item sets. The thirty additional "filler" items included in the field test to accommodate test book design constraints were excluded from the analyses. There were 13, 4, and 13 filler items, respectively, in Forms A, B, and C.

Table 7

Raw Score Summary Statistics for CAP and Stanford Items Combined: Form A (n = 1138)

Score	Number of Items	M	SD	Percent Correct
Subtests:				
Concepts of Number	39	28.1	7.4	72.1
Mathematics Computation	38	24.0	7.7	63.0
Mathematics Applications: Part 1	37	26.7	7.0	72.2
Mathematics Applications: Part 2	26	16.1	5.7	61.8
Total Mathematics Applications	63	43.8	11.8	67.9
Total Test	140	94.9	24.1	67.8
Skill Areas:				
I.A.1.	22	16.2	4.2	73.6
III.A.3.	26	18.2	5.4	69.8
III.A.5.	12	5.8	3.3	48.2
III.A.7.	17	11.9	3.7	70.1
V.A.2.	16	11.7	3.3	73.1
VI.B.	26	16.1	5.7	61.8
IX.A.	21	15.0	4.4	71.4

Table 8

Raw Score Summary Statistics for CAP and Stanford Items Combined: Form B (n = 1145)

Score	Number of Items	M	SD	Percent Correct
Subtests:				
Concepts of Number	40	29.1	7.4	72.8
Mathematics Computation	45	30.1	9.1	66.9
Mathematics Applications: Part 1	38	26.1	7.6	68.6
Mathematics Applications: Part 2	26	15.3	5.6	58.8
Total Mathematics Applications	64	41.4	12.4	64.6
Total Test	149	100.6	26.4	67.5
Skill Areas:				
I.A.2.	16	12.2	3.5	76.3
II.A.1.	24	16.9	4.6	70.5
III.A.2.	20	15.3	3.8	76.7
III.A.6.	25	14.8	6.3	59.0
III.B.1.	17	12.4	3.6	73.0
III.B.2.	26	15.3	5.6	58.8
VIII.B.	21	13.7	4.6	65.1

Table 9

Raw Score Summary Statistics for CAP and Stanford Items Combined: Form C (n = 1187)

Score	Number of Items	M	SD	Percent Correct
Subtests:				
Concepts of Number	37	23.4	6.9	63.3
Mathematics Computation	45	32.7	9.2	72.7
Mathematics Applications: Part 1	33	23.1	6.0	70.0
Mathematics Applications: Part 2	25	15.4	5.4	61.4
Total Mathematics Applications	58	38.4	10.7	66.2
Total Test	140	94.6	23.8	67.6
Skill Areas:				
II.A.2.	25	15.8	4.6	63.1
III.A.1.	26	20.0	5.1	77.1
III.A.4.	19	12.7	4.7	66.7
III.B.3.	25	15.4	5.4	61.4
IV.A.2.	12	7.6	2.9	63.6
V.A.1.	12	8.2	2.5	68.3
VIII.A.	21	14.9	4.1	70.8

The traditional item statistics obtained from the field test are summarized in Tables 10 and 11. Table 10 presents the number of items in the set, the mean p-value for the set, and its standard deviation. The same information is presented in Table 11 for biserial correlations.

Table 10

Mean P-Values and Standard Deviations For CAP and Stanford Item Sets

Skill Area		Number of Items	1981-82 M	1981-82 SD	1985 M	1985 SD
1. I.A.1.	CAP	15	.67	.18	.72	.16
	Stanford	7	.68	.13	.77	.12
2. I.A.2.	CAP	10	.65	.13	.75	.08
	Stanford	6	.72	.14	.78	.10
3. II.A.1.	CAP	15	.67	.17	.70	.15
	Stanford	9	.66	.18	.72	.14
4. II.A.2.	CAP	19	.57	.23	.61	.22
	Stanford	6	.61	.14	.69	.14
5. III.A.1.	CAP	13	.78	.16	.73	.13
	Stanford	13	.77	.08	.73	.07
6. III.A.2.	CAP	12	.78	.15	.72	.13
	Stanford	8	.77	.09	.83	.07
7. III.A.3.	CAP	14	.71	.12	.62	.12
	Stanford	12	.74	.15	.80	.14
8. III.A.4.	CAP	13	.58	.21	.61	.13
	Stanford	6	.67	.05	.79	.06
9. III.A.5.	CAP	12	.54	.16	.49	.11
	Stanford	0	--	--	--	--
10. III.A.6.	CAP	14	.54	.14	.52	.13
	Stanford	11	.54	.15	.68	.13
11. III.A.7.	CAP	11	.65	.18	.68	.18
	Stanford	6	.62	.12	.73	.14
12. III.B.1.	CAP	12	.68	.14	.68	.11
	Stanford	5	.80	.10	.85	.08
13. III.B.2.	CAP	20	.53	.15	.54	.15
	Stanford	6	.72	.15	.75	.10
14. III.B.3.	CAP	15	.48	.16	.54	.14
	Stanford	10	.62	.07	.73	.09
15. IV.A.2.	CAP	12	.58	.10	.64	.12
	Stanford	0	--	--	--	--

(Con'd)

Table 10 (Con'd)

Skill Area		Number of Items	1981-82 M	1981-82 SD	1985 M	1985 SD
16. V.A.1	CAP	12	.66	.14	.69	.17
	Stanford	0	--	--	--	--
17. V.A.2.	CAP	12	.61	.20	.70	.14
	Stanford	4	.62	.05	.83	.03
18. VI.B.	CAP	16	.58	.14	.60	.15
	Stanford	10	.58	.13	.65	.11
19. VIII.A.	CAP	15	.69	.15	.71	.14
	Stanford	6	.62	.12	.70	.11
20. VIII.B.	CAP	15	.65	.12	.61	.11
	Stanford	6	.65	.13	.74	.15
21. IX.A.	CAP	15	.70	.13	.71	.14
	Stanford	6	.62	.10	.73	.11
Overall:						
	CAP		.63	.07	.65	.08
	Stanford		.67	.07	.75	.06

Table 11

Mean Biserial Correlations and Standard Deviations For CAP and Stanford Item Sets

Skill Area		Number of Items	1981-82 M	1981-82 SD	1985 M	1985 SD
1. I.A.1.	CAP	15	--	--	.53	.10
	Stanford	7	.54	.07	.49	.06
2. I.A.2.	CAP	10	--	--	.63	.07
	Stanford	6	.59	.08	.58	.07
3. II.A.1	CAP	15	--	--	.43	.15
	Stanford	9	.58	.06	.55	.09
4. II.A.2.	CAP	19	--	--	.44	.14
	Stanford	6	.57	.11	.57	.07
5. III.A.1.	CAP	13	--	--	.65	.20
	Stanford	13	.53	.12	.58	.13
6. III.A.2.	CAP	12	--	--	.59	.24
	Stanford	8	.63	.04	.52	.07
7. III.A.3.	CAP	14	--	--	.59	.13
	Stanford	12	.71	.12	.59	.08
8. III.A.4.	CAP	13	--	--	.68	.06
	Stanford	6	.65	.09	.54	.05
9. III.A.5.	CAP	12	--	--	.56	.08
	Stanford	0	--	--	--	--
10. III.A.6.	CAP	14	--	--	.62	.10
	Stanford	11	.62	.04	.62	.13
11. III.A.7.	CAP	11	--	--	.56	.18
	Stanford	6	.66	.07	.68	.09
12. III.B.1.	CAP	12	--	--	.58	.19
	Stanford	5	.66	.02	.56	.04
13. III.B.2.	CAP	20	--	--	.49	.12
	Stanford	6	.60	.12	.56	.06
14. III.B.3.	CAP	15	--	--	.52	.12
	Stanford	10	.62	.06	.54	.09
15. IV.A.2.	CAP	12	--	--	.51	.09
	Stanford	0	--	--	--	--

(Con'd)

Table 11 (Cont)

Skill Area		Number of Items	1981-82 M	1981-82 SD	1985 M	1985 SD
16. V.A.1.	CAP	12	--	--	.48	.12
	Stanford	0	--	--	--	--
17. V.A.2.	CAP	12	--	--	.55	.14
	Stanford	0	.49	.06	.59	.10
18. VI.B.	CAP	16	--	--	.53	.09
	Stanford	10	.56	.10	.53	.12
19. VIII.A.	CAP	15	--	--	.51	.13
	Stanford	6	.58	.04	.49	.09
20. VIII.B.	CAP	15	--	--	.52	.10
	Stanford	6	.63	.09	.55	.12
21. IX.A.	CAP	15	--	--	.56	.12
	Stanford	6	.59	.09	.56	.08
Overall:						
	CAP		--	--	.55	.07
	Stanford		.60	.05	.56	.04

Two-Parameter Item Response Theory Analyses

The item response data from the field test were analyzed using the two-parameter logistic ogive model (Bock & Mislevy, 1981; Bock, Mislevy, & Woodson, 1982). Each of the 21 skill areas in each field test form was analyzed twice, once using the combined CAP and Stanford items and once using only the CAP items. A sample of 1000 students was randomly selected for the analyses performed on each form.

BILOG, a computer program that performs item analysis and test scoring based on the two-parameter model, was used to perform the analyses (Mislevy & Bock, 1984). The same model and estimation procedures were used in all calibrations.

In this specific application, the distribution of scale scores in each of the samples was assumed to be approximately normal. That is, a normal prior was assumed for the student scale score distribution. The mean and variance of this distribution were constrained to be "0" and "1," respectively, to resolve indeterminancies of scale and origin. Ten quadrature points and weights were used to approximate integration over this normal density.

To enhance the stability of the item parameter estimates, prior distributions were placed on the item parameters. (Note: When prior distributions are placed on the item parameters, Bayes modal rather than maximum likelihood estimates of the parameters are obtained.) In each cycle of the estimation procedure that generates the estimates, the mean of each of these priors was updated to reflect the average location of the corresponding provisional item parameter estimates. Placement of prior distributions on item parameter estimates essentially restricts the range of possible estimates to "reasonable" values. Employing this procedure avoids the problem of infinite slopes, for example. For further information on the use of prior distributions, see the BILOG manual (p.15-26).

During the item calibration, a maximum of ten EM estimation cycles and a convergence criterion of 0.01 were used. The EM cycles (Dempster, Laird & Rubin, 1977) were followed by a Newton-Raphson cycle to obtain standard errors for the estimates.

Indices of item fit were generated following the final estimation cycle. The method used to derive these approximate chi-squares is described in the BILOG manual (p.4). An overall measure of model fit was also generated for each skill area. An approximate chi-square was computed by summing the individual chi-square indices for the items in the skill area. The degrees of freedom for this overall index is equal to the sum of those of the individual items.

Table 12

Means and Standard Deviations of the Item Parameter Estimates and Measures of Fit Obtained from the Two-Parameter Analysis of the Combined CAP and Stanford Items by Skill Area

Skill Area		Items	B-Value M	B-Value SD	A-Value M	A-Value SD	Percent Misfit
1. I.A.1.	CAP	17	-1.15	.87	.78	.19	11.8
	Stanford	7	-1.50	.79	.99	.12	14.3
2. I.A.2.	CAP	10	-1.10	.44	.67	.16	20.0
	Stanford	6	-1.29	.55	.73	.15	33.3
3. II.A.1.	CAP	15	-1.06	.83	1.07	.71	26.7
	Stanford	9	-1.06	.73	.83	.20	33.3
4. II.A.2.	CAP	19	- .46	1.42	1.15	.55	26.3
	Stanford	6	- .77	.52	.71	.17	16.7
5. III.A.1	CAP	13	- .93	.54	.79	.35	7.7
	Stanford	13	-1.72	.81	.87	.27	0.0
6. III.A.2.	CAP	12	- .93	.51	1.06	1.15	0.0
	Stanford	8	-1.90	.66	.96	.15	0.0
7. III.A.3.	CAP	14	- .44	.57	.80	.33	14.3
	Stanford	12	-1.47	.78	.72	.19	25.0
8. III.A.4.	CAP	13	- .44	.48	.56	.10	76.9
	Stanford	6	-1.45	.35	.87	.12	50.0
9. III.A.5.	CAP	12	.07	.46	.74	.20	8.3
	Stanford	0	--	--	--	--	---
10. III.A.6	CAP	14	- .11	.64	.78	.21	21.4
	Stanford	11	- .85	.71	.70	.27	18.2
11. III.A.7.	CAP	11	- .82	.76	.84	.44	9.1
	Stanford	6	- .86	.51	.55	.13	0.0
12. III.B.1.	CAP	12	- .75	.50	.83	.43	8.3
	Stanford	5	-1.85	.70	.77	.08	20.0
13. III.B.2.	CAP	20	.13	1.32	1.06	.69	20.0
	Stanford	6	-1.13	.46	.71	.10	66.7
14. III.B.3.	CAP	15	- .11	.67	.93	.28	20.0
	Stanford	10	-1.06	.39	.85	.24	0.0

(Con'd)

Table 12 (Con'd)

Skill Area		Items	B-Value M	B-Value SD	A-Value M	A-Value SD	Percent Misfit
15. IV.A.2.	CAP	12	- .60	.68	.83	.30	25.0
	Stanford	0	--	--	--	--	---
16. V.A.1	CAP	12	- .86	.84	.90	.31	16.7
	Stanford	0	--	--	--	--	---
17. V.A.2.	CAP	12	- .87	.57	.83	.33	8.3
	Stanford	4	-1.55	.36	.70	.20	0.0
18. VI.B.	CAP	16	- .47	.77	.91	.25	18.8
	Stanford	10	- .70	.65	.92	.32	0.0
19. VIII.A.	CAP	15	-1.10	.83	.95	.34	13.3
	Stanford	6	-1.07	.77	.97	.23	0.0
20. VIII.B.	CAP	15	- .53	.54	.91	.27	33.3
	Stanford	6	-1.11	.72	.84	.32	16.7
21. IX.A.	CAP	15	- .98	.76	.81	.26	26.7
	Stanford	6	-1.04	.54	.80	.15	16.7
Overall:							
	CAP		- .64		.87		20.1
	Stanford		-1.24		.80		16.1

Analysis of CAP and Stanford Items Combined. Table 12 summarizes the results of the two-parameter IRT analyses. Here means and standard deviations are reported for each of the item parameter estimates. To facilitate comparisons, means and standard deviations are reported separately for CAP and Stanford items even though these items were calibrated as a single set of items.

Phase 3: Identification of the Multi-Purpose Item Set

One of the goals of this study was to examine the concept of multi-purpose testing as a methodology for consolidating and unifying California's current testing and assessment system. The third phase of the study consisted of identifying a "core" set of multi-purpose Stanford items from among those included in the field test. This item set could then be used to generate the kind of information needed by the state, as well as norm-referenced information required by districts. The selection process used to identify these items is described next.

1. Selection Criteria. Two criteria were used to select the final set of multi-purpose Stanford items. The first criterion was commonality of skills. The items selected must measure a skill area, or specific subskill within a skill area, also measured by the California Assessment Program. All of the Stanford items in the field test satisfied this criterion based on the content review and item mapping conducted during Phase 1. The final set would also have to satisfy this criterion empirically.

The second selection criterion was representativeness. The final set of items must reflect the objective structure of the Stanford as described in Table 3. The set must include a proportionate number of items from each objective and cluster in the Stanford since these items would be used to estimate Stanford mathematics subtest and total test scores.

2. Selection Target. The target number of Stanford items to be selected was set at 54: 16 items from the Concepts of Number subtest, 20 from the Mathematics Computation subtest, and 18 items from the Mathematics Applications subtest. The target figure was based on the fact that 15 to 20 items is generally an acceptable number from which to estimate a score. Further, this number would include about half the items in each subtest and would allow some items from every objective in Stanford to be represented.

More than twice the number of items needed to meet the target, namely 137, had therefore been included in the field test. The pool of potential items included 45 Concepts of Number items, 51 Mathematics Computations items, and 41 Mathematics Applications items.

3. Selection Process

The materials prepared for the item mapping conducted as part of Phase 1 of the study were used again during the final selection of the multi-purpose item set. The text and selected item statistics for all 137 Stanford items under consideration had been pasted on pages to facilitate comparisons between CAP and Stanford items purported to measure the same thing. Additional item statistics based on the results of the field test were then added to the item information already displayed. Both traditional and IRT statistics were included. The items, both CAP and Stanford, were then organized by skill area so that all the items measuring the same skill could be viewed together.

The chi-square index of fit generated for each Stanford item was used to satisfy the first selection criterion. For each skill area under study, the values from the combined analysis of the CAP and Stanford items measuring that skill area were reviewed. All Stanford items with chi-square fit indices greater than two times their degrees of freedom were eliminated from the pool. A total of 21 items were removed based on their fit indices. Eight of these were Concepts of Number items, 6 were Mathematics Computation items, and 7 were Mathematics Applications items.

The pages of item text and statistics for the Stanford-only items were then reorganized around the objective structure of the Stanford so that all items measuring the same Stanford objective could be viewed together. Twenty-two of these items were from Form F of the Stanford. To further reduce the size of the pool, the decision to consider only items from Form E was made. The remaining items measuring each objective were then reviewed on an item-by-item

basis, and the "best" overall items were selected. The best items were defined to be those with small chi-square fit indices, large biserial correlations, and p-values within the range of the skill element in which they had been included. A total of 51 items were identified.

Three additional items, not included in the field test, were then added to the core set, making the final total 54 items. These items were added in order to reflect more accurately the objective structure of Stanford. This was necessary because three objectives measured in Stanford were not also included in the California's core curriculum at this grade level. The first objective was in the Concepts of Number subtest: Interpret integers. This objective was measured by two items in Stanford, both of which involved negative numbers. The concept of negative numbers is not evaluated at Grade 6 in California.

The second objective was from the Mathematics Computation subtest: Estimate sums, products and quotients of whole numbers. Three items measure this objective in Stanford. While some of the Grade 6 CAP items involve estimation, no really good match could be found for these Stanford items.

The third objective was from the Mathematics Applications subtest: Express problems as number sentences. Three Stanford items measured this objective. Again, however, a good match of these items to a particular CAP skill area could not be found.

4. The Final Set of Multi-Purpose Items

The final set of multi-purpose items contained 54 Stanford items. All items were from the Intermediate 2 level of the series, Form E. Sixteen items were from the Concepts of Number subtest, 20 items were from the Mathematics Computation subtest, and 18 items were from the Mathematics Applications subtest. The objective structure

of Stanford at the Intermediate 2 level is presented in Table 13. Also reported there is the total number of items in the Stanford and the number of these in the final multi-purpose item set.

Table 14 presented individual item information about the set of multi-purpose items. The items are presented in the order in which they appear in their respective Stanford subtests. Item p-values, biserial correlations, b-values, a-values and chi-square fit indices are reported for each item. These statistics are based on the analyses performed following the field test on the combined CAP and Stanford items measuring each skill element. Also reported is the skill area they measure.

Table 14

Selected Item Statistics for the 54 Stanford Items in the Multi-Purpose Item Set

Item ID	P-Value	Biserial Correlation	B-Value	A-Value	Chi Square	df	p	Skill Area
Concepts of Number:								
1. CN3	.63	.51	-.64	.91	7.0	7	.430	I.A.1
2. CN4	.82	.49	-1.72	.93	5.4	7	.609	I.A.1.
3. CN6	.69	.55	-.85	.86	10.8	6	.094	I.A.2.
4. CN8	.91	.61	-2.07	.68	8.9	5	.112	II.A.1.
5. CN11	.82	.51	-1.67	.91	1.8	6	.935	I.A.2
6. CN13	.60	.43	-.51	1.07	4.3	8	.826	I.A.1
7. CN15	.73	.70	-.82	.50	0.5	5	.991	III.A.7.
8. CN17	.49	.52	.03	.79	3.7	7	.812	III.A.7.
9. CN19	.75	.55	-1.14	.81	1.6	7	.978	II.A.1.
10. CN21	.76	.51	-1.18	.77	10.5	6	.104	I.A.2.
11. CN24	.73	.68	-.83	.53	11.9	6	.064	III.A.7.
12. CN25	.78	.64	-1.16	.68	9.2	7	.237	II.A.1.
13.* CN27	None							
14. CN30	.79	.55	-1.45	.86	9.1	7	.246	II.A.1.
15. CN31	.76	.39	-1.64	1.22	5.0	8	.754	II.A.1.
16. CN33	.87	.67	-1.43	.46	7.8	4	.099	II.A.2.
M:	.74	.55	-1.14	.80				
SD:	.11	.09	.55	.21				
Mathematics Computation:								
1. MC2	.80	.48	-1.77	1.09	8.8	8	.360	III.A.1.
2. MC3	.73	.55	-1.07	.90	7.1	8	.530	III.A.1.
3. MC6	.81	.73	-1.28	.64	11.3	7	.127	III.A.1.
4. MC8	.77	.66	-1.14	.73	2.8	7	.899	III.A.1.
5. MC11	.77	.56	-1.28	.86	3.0	6	.808	III.A.2.
6. MC12	.74	.45	-1.31	1.06	5.0	7	.657	III.A.2.
7. MC14	.89	.41	-2.91	1.23	3.0	6	.806	III.A.2.
8. * MC16	None							
9. MC18	.95	.65	-2.32	.54	2.5	2	.285	III.A.3.
10. MC22	.65	.49	-.78	1.03	2.8	9	.972	III.A.3.
11. MC23	.65	.52	-.74	.91	4.9	8	.774	III.A.3.
12. MC26	.57	.47	-.35	1.07	4.3	8	.831	III.A.3.
13. MC28	.86	.57	-1.74	.76	11.4	6	.075	III.A.4.
14. MC31	.74	.56	-1.13	.86	14.7	8	.065	III.A.4.
15. MC33	.77	.51	-1.36	.92	26.2	7	.491	III.A.4.
16. MC36	.52	.75	-.06	.52	7.9	7	.339	III.A.6.
17. MC37	.58	.80	-.29	.43	3.8	6	.709	III.A.6.
18. MC40	.68	.61	-.77	.78	5.9	7	.553	III.A.6.
19. MC41	.62	.66	-.47	.70	4.2	8	.840	III.A.6.
20. MC43	.73	.58	-1.03	.81	8.1	7	.322	III.A.6.
M:	.73	.58	1.15	.83				
SD:	.11	.11	.70	.21				

(Con'd)

Table 14 (Con'd)

Item ID	P-Value	Biserial Correlation	B-Value	A-Value	Chi Square	df	p	Skill Area
Mathematics Applications:								
1. MA2	.75	.61	-1.05	.67	3.9	7	.790	VI.B.
2. MA4	.72	.61	-.89	.70	10.7	7	.149	VI.B.
3. MA5	.70	.53	-.90	.82	14.8	8	.630	III.B.3.
4. MA8	.84	.53	-1.68	.80	7.4	7	.389	III.B.3.
5. MA12	.63	.62	-.51	.66	6.9	8	.549	III.B.3.
6. MA14	.59	.36	-.59	1.41	11.4	9	.247	III.B.3.
7. MA15	.62	.64	-.47	.64	5.4	8	.721	III.B.2.
8. MA16	.63	.71	-.47	.53	5.1	7	.644	VI.B.
9. * MA18	None							
10. MA21	.68	.58	-.75	.76	6.4	8	.603	IX.A.
11. MA25	.88	.44	-2.51	1.07	8.7	7	.273	VIII.A.
12. MA27	.62	.39	-.71	1.28	3.0	8	.932	VIII.B.
13. MA30	.80	.63	-1.23	.59	2.9	6	.820	V.A.2.
14. MA32	.63	.58	-.54	.75	5.5	8	.701	VI.B.
15. MA35	.77	.42	-1.56	1.11	8.2	9	.512	VI.B.
16. MA37	.83	.56	-1.56	.77	11.8	7	.106	III.B.3.
17. MA38	.69	.45	-1.01	1.06	9.6	9	.386	III.B.3.
18. MA39	.44	.39	.34	1.34	9.9	9	.429	VI.B.
M:	.69	.53	-.95	.88				
SD:	.11	.10	.64	.28				
Overall:								
M:	.72	.56	-1.08	.84				
SD:	.11	.10	.63	.23				

* Additional items not in the field test.

Phase 4: Validation Study

The fourth and final phase was a validation study. It examined the accuracy with which norm-referenced test information could be estimated from the set of multi-purpose items identified in Phase 3.

The general strategy employed in the validation study was to estimate Stanford total test and subtest scores in mathematics by applying item response theory (IRT) techniques to the multi-purpose item set identified in Phase 3. These "estimated" Stanford scores were then compared with the total test and subtest Stanford scores actually earned by students. The accuracy of this approach was evaluated using pairs of estimated and actual scores.

Validation Sample

The validation sample consisted of 1,061 Grade 6 students from two California school districts. The Stanford mathematics item response data, collected during each district's regular spring testing program, were used in the study. The item response data from 1,031 of the 1,061 students were used in the validation study. Data were included in the analyses only if the students took all three of the Stanford mathematics subtests and attempted at least three of the first six items in each.

The Stanford Scaled Score Scale

Like the California Assessment Program, the measurement model used to develop, equate, and scale the Stanford is based on item response theory (IRT). Specifically, the one-parameter or Rasch (Rasch, 1960) model was employed.

One of the features of the Stanford is that all levels and forms are linked together by a single scaled score system. Any raw score earned by a student can therefore be converted to its scaled-score

equivalent, making comparisons across levels and forms of the test possible on this common scale.

IRT played a central role in the development of the Stanford scaled score system. Initially, the Rasch model was used to place the raw scores on all ten levels in the series on the same scale. An equating study was conducted in which adjacent levels of the test were administered to the same group of students. The items in pairs of adjacent levels were then calibrated together and mean Rasch item difficulty estimates computed by level. These values served as equating constants to convert the item difficulty estimates of one test level to the scale of the next test level. The appropriate equating constants were then added to the item difficulty estimates from the fall 1982 standardization to produce an equated Rasch ability scale.

The final scaled score scale was a "convenience scaling" of this Rasch ability scale. It was developed by applying a linear transformation to the scale, thus converting it to a three-digit integer scale with only positive integers.

Estimated Test and Subtest Scores

The Rasch model was used to generate estimated Stanford total test and subtest scaled scores in mathematics. The same process, described next, was employed for each of the three subtests and for the total test. It consisted of applying the Rasch item difficulty estimates for the multi-purpose item set to the item response data of students in the validation sample. The particular set of Rasch item difficulty estimates applied were those obtained in the equating study from which the Stanford scaled score scale was developed. These Rasch difficulty estimates had been adjusted by the appropriate equating constant to reproduce the Rasch ability scale underlying the Stanford. This scale was then transformed into the scaled score scale using the same linear transformation.

Table 16 presents the Pearson product-moment correlations between pairs of estimated and actual scaled scores. Also reported there is the scaled root mean square difference (RMSQ) between pairs of estimated and actual scaled scores. The RMSQ is an index of the overall difference between the two sets of scaled scores, like an empirical standard error. The RMSQ values reported in Table 16 are scaled by the variances of the scores to permit comparisons across subtests.

Table 16

Pearson Product-Moment Correlations and Scaled Root Mean Square Differences for Pairs of Estimated and Actual Scaled Scores

Subtest	Correlation	RMSQ
Concepts of Number	.937	.362
Mathematics Computation	.956	.302
Mathematics Applications	.945	.337
Total Mathematics	.976	.221

Finally, pairs of estimated and actual scaled scores were compared in terms of their standard errors of measurement (SEM). As would be expected, the SEMs for the tests with the full compliment of items were always lower than the SEMs for the corresponding shorter tests. However, the size of the decrease is of interest since it represents a change in the precision of the scores. In the mid-range of the distribution, the total test SEM decreased roughly three scaled-score points, from 10 to 7, while those for the subtests decreased roughly 5 scaled-score points, from 17 or 18 to 12 or 13.

Discussion

The purpose of the study reported here was to examine the concept of multi-purpose testing. A discussion of the results will be organized around the four major phases of the study. The four phases are:

Phase 1: Content Review
Phase 2: Field Test
Phase 3: Identification of a Multi-Purpose Item Set
Phase 4: Validation Study.

1. Phase 1: Content Review. The goal of the Comprehensive Assessment System is to provide a consolidated testing system for the state of California. One of the features of the system is a reduction in the actual number of test items taken each year by California students. This could be accomplished by eliminating duplicate testing of skills and by carefully designing tests to yield information for more than one purpose.

A basic premise of such a testing system is that the various testing programs currently conducted in California actually have in common certain essential skills or objectives. This premise was the focus of Phase 1 of the study.

Content Review. A indepth content review was performed for two testing programs: the California Assessment Program and the Stanford Achievement Test, Seventh Edition. One content domain, mathematics, and one grade level, Grade 6, were considered.

The results of the review and subsequent item mapping are reported in Table 4. Given the differences in purpose and therefore specificity of the two testing programs, the results are not surprising. The California Assessment Program measures 33 mathematics skill areas at Grade 6, each of which is assessed by between 10 and 25 items. The entire domain includes some 480 items,

all of which are administered each year. One of the strengths of this program lies in its systematic coverage of, and reporting about, very detailed components of California's educational program. This degree of specificity is achieved through the use of multiple-matrix sampling and item response theory at the cost of information about individual students.

The Stanford is also broad based. However, the mathematics subtests appropriate for Grade 6 include only fourteen "skill areas" or objectives, each measured by from three and 22 items. The entire test has only 117 items. The purpose of the test, and the series in general, is to assess student achievement in the basic skills areas, and the series was designed to reflect what is being taught in schools throughout the country.

The major difference between the skills assessed by these two testing programs is one of breadth or scope. The California Assessment Program includes many more skills and many more items measuring those skills. The results of the item mapping clearly indicate that almost all of the mathematics skills assessed by the Stanford are shared with the California Assessment Program. Again however, the latter assesses more skills, and subskills, and assesses them with more items.

It is interesting to note that although the two programs differed significantly in breadth, the specificity of the items themselves is quite similar across programs. This fact is encouraging since it suggests that one set of items could be used to measure those skills common to both programs.

There were four skill areas assessed by the California Assessment Program that were not assessed in the Stanford. They were 1) Measurement Skills including Metric Units, US Customary Units, and Length, Area and Volume, 2) Probability and Statistics, 3) Graphs and Functions, and 4) application of skills in Number Theory and

Geometry. In general, the California Assessment Program has a greater focus on the application of skills.

In summary, then, the results of the content review and item mapping suggest that, for the grade level and content area studied, at least two of California's testing programs assess common skills. This result supports the concept of a consolidated testing system such as the Comprehensive Assessment System. An indepth review of content and skills at other grade levels and content areas would, of course, be required to confirm such a generalization. However, the fact that the basic structure of the skill areas assessed at other grade levels and in other content areas is quite similar to the one studied here suggests that extending the concept of multi-purpose testing beyond Grade 6 mathematics will be feasible.

2. Phase 2: Field Test. Phase 2 of the study consisted of a mathematics field test conducted at Grade 6. Six school districts were represented from central and southern California. Some 4,000 California students in over 60 different schools participated in the testing.

Field Test Sample. The methodology of the study did not require that the field test sample be a representative sample of students from the state of California. It did, however, require a sample that adequately represented all levels of mathematics ability. Accordingly, school districts were instructed to administer the field test to all schools in their district. Where this was not possible, school districts were asked to select schools that would represent a wide range of abilities.

All three forms of the field test were administered in every district. The strategy for assigning field test forms to classrooms within districts was designed to ensure that the students taking each form were as comparable as possible with respect to mathematics ability. Classrooms within districts were therefore ranked from

lowest to highest based on the mean total mathematics scale score earned by their school. Forms were then randomly assigned to classrooms. This procedure resulted in a more or less equal number of students from each district taking each field test form. About half this group was male and half was female.

The mean 1984 CAP mathematics scale score calculated for eacn form of the field test suggests that the sample was above average. Values of 274.4, 269.3, and 271.2, respectively, for Forms A, B, and C were obtained. The mathematics scaled-score scale used by the California Assessment Program has a mean of approximately 250 and a standard deviation of 50. It should be noted, however, that the mean values are estimates of the mathematics ability of participating districts. They are based on school-level data applied at the classroom level, and were used merely as a guide in assigning field test forms.

Additional evidence in support of the conclusion that the field test sample was above average can also be cited. The average percent correct for 18 of the 21 skill areas studied was higher for this administration (1985; values not reported in this paper) than for the 1981-1982 administration of the same item sets.

A similar improvement in performance since 1981-1982 was also observed for overall mean item difficulty. While some of these differences can certainly be attributed to improved educational practices in the state of California, it seems likely that some of the differences can be attributed to the above average ability level of the sample.

Although the sample appeared to be generally above average, the ability level of the students taking each field test form appears to be quite comparable across forms. This is evidenced by the overall percent correct for each form reported in Tables 7 through 9 for Forms A, B, and C, respectively. The average percents correct were

67.8, 67.5, and 67.6, respectively. These values are consistent with those reported in Table 6, where the average difficulty of each form was computed using p-values from previous administrations of the items. These values were .64, .65, and .64, respectively. The increase in mean p-values also suggests that the field test sample was above average.

Comparison of CAP and Stanford Item Sets. The purpose of the field test was to collect empirical information about CAP and Stanford items measuring the same skills. These comparisons will be discussed in terms of traditional and IRT information generated from the analysis of the field test data.

Overall, the sets of items compared favorably. In general, the Stanford items appeared to be somewhat easier, although both sets were comparable with respect to discrimination. The evidence supporting these conclusions is reviewed next.

Difficulty. In general, the Stanford items included in the field test appeared to be easier than their CAP counterparts measuring the same skill area. The mean p-values obtained in this study are summarized in Table 11 for CAP and Stanford items. Almost without exception, the mean value for the Stanford items measuring a given skill area is larger than the mean for the CAP items measuring that same skill area. The overall mean p-values reflect this phenomenon as well. These values were .65 and .75, respectively, for the CAP and Stanford item sets.

Some of the difference in item difficulty is the result of the item sequencing strategy used to assign items to positions within field test forms. In most instances, the Stanford items appeared before their CAP counterparts for a given skill area. This strategy was necessary because of differences between CAP and Stanford items in terms of numbers of answer options. Where CAP and Stanford items were intermixed, as in the Concepts of Number subtests, the differeces in mean p-values, while in the same direction, were not as great.

The fact that the Stanford items were generally easier than CAP items measuring the same skill area is substantiated by results obtained from the two-parameter IRT analyses performed on the field test data. The mean item difficulty estimates (b-values) for CAP and Stanford items are summarized in Table 13. Mean values are reported there for each item set although the two sets were actually analyzed together. The overall mean b-values reported there were -.64 and -1.24, respectively. The b-value scale ranges roughly from -4.00 to +4.00, with negative values indicating easier items.

Although the Stanford items were generally easier than their CAP counterparts, the two item sets were more or less equally discriminating. The overall mean biserial correlations reported in Table 12 were .54 and .56, respectively, for CAP and Stanford items. Further evidence in support of this observation can also be found in Table 14, where the results for the two-parameter analyses are summarized. The overall average mean slope estimates (a-values) were .87 and .80, respectively, for the CAP and Stanford item sets.

Model Data Fit. A final comparison between CAP and Stanford items can be made based on their fit indices. In general, both sets of items fit the IRT model employed quite well. Using chi-square values less than two times their degrees of freedom as the criterion for fitting items, only 59 of the 292 CAP items, or approximately 20 percent, did not fit the two-parameter model. Only 22 of the 137 Stanford items, or 16 percent, did not fit. It should be noted that these results are based on the combined analysis of sets of CAP and Stanford items measuring the same skill area. The fact that both sets of items seemed to fit the two-parameter model equally well is very encouraging, since it further supports the feasibility of multi-purpose testing.

While the overall fit of CAP and Stanford items was good in general, there were two exceptions. First, the fit of both the CAP and Stanford items measuring skill area III.A.4., "Addition and Subtraction of Decimals," was poor, with 76.9 and 50 percent of the items, respectively, misfitting. Second, the fit of Stanford items measuring skill area III.B.2., "One-step Operations involving Whole Numbers," was also poor, with 66.7 percent of the items misfitting. An examination of the items in these two skill areas did not reveal the reason for the large percent of misfitting items. It is hypothesized that lack of unidimensionality may be the source of poor fit observed in these two instances.

Conclusions. In summary, then, the results of the field test supported the conclusions reached by the Phase 1 content review and item mapping discussed earlier. The data showed that the Stanford items purported to measure skills also assessed by the California Assessment Program did in fact behave very much like their CAP counterparts. In general, they were easier than the CAP items, but they were equally discriminating and fit the two-parameter model as well as the corresponding CAP items measuring the same skill.

3. Phase 3: Identification of Multi-Purpose Item Set. The third phase of the study consisted of identifying a core set of multi-purpose items from among those included in the field test. The first selection criterion, referred to here as commonality of skills, was that the items must measure a skill area, or subskill within a skill area, also measured by the California Assessment Program. The second criterion was representativeness. As a set, the items selected must reflect the objective structure of the Stanford.

Commonality. A total of 54 Stanford items was selected. In general, the multi-purpose items in this set were more like their CAP counterparts, in terms of their overall statistical qualities, than the total Stanford item set from which they were selected. The

items in the core set were still easier than those in the CAP item set. The overall mean p-value reported in Table 14 for the multi-purpose item set was .72. This value, however, is somewhat lower than the mean value for the entire pool of items from which they were selected. The overall mean p-value for the entire pool, reported in Table 10, was .75. The overall mean p-value for the CAP items was .65.

Like the CAP and Stanford item sets discussed earlier, the core set of multi-purpose items had fairly large discrimination indices as measured by their biserial correlations. The mean biserial correlations for the CAP items, the total Stanford pool, and the core set of multi-purpose items were .55, .56, and .56, respectively.

The results of the two-parameter IRT analyses were consistent with those reported for the traditional item statistics. The mean item difficult estimate for the multi-purpose item set, reported in Table 14, was -1.08. This value fell between the overall mean for the CAP items (-.64) and the overall mean for the Stanford pool (-1.24). These values were reported in Table 12. The overall mean item slope estimates for the three item sets were quite similar.

Representativeness. The second selection criterion was that the final core set must, as a set, reflect the objective structure of the Stanford. This criterion was also satisfied by the set of items finally selected. Table 13 presents the distribution of multi-purpose items across the objectives assessed in Stanford. All objectives within the structure are represented in the core set. In order to meet this criteria, three additional items not included in the field test were added to the set.

It should be noted that during the selection process the "best" Stanford items were identified. However, many of the items not selected could have been substituted for those that were finally selected since there were many "good" items in the pool. This fact

is important because it expands the range of skill areas that might be assessed in any given administration. The implication here is that different sets of multi-purpose items could be identified, each measuring a different configuration of skill areas. This fact has implications for test security also.

4. Phase 4: The Validation Study. The fourth and final phase of the study examined the accuracy of norm-referenced test information generated from a core set of multi-purpose items. Rasch item difficulty estimates for the multi-purpose item set were used to generate four Stanford mathematics scaled scores for a sample of about 1,000 California students. The Rasch difficulty estimates used were from the equating and standardization of the Stanford conducted in 1982. The estimated scaled scores were then compared to those actually earned by the students.

In general, the estimated scaled scores compared quite favorably to those actually earned by students. The correlations between pairs of scores were large, and the scaled root mean squared differences were small. As was anticipated, some decrease in the accuracy of these scores was observed due to the decreased number of items being used to generate them. The decrease was, however, within acceptable limits. A discussion of the validation sample and the details of these conclusions just summarized will be presented next.

Validation Sample. The validation sample consisted of over 1,000 students from two California school districts. The sample included all the sixth graders tested in each district during its spring 1985 testing program. During this testing, students took the complete Stanford battery appropriate for their grade placement. Only their mathematics data were considered in this study.

The mean scaled scores actually earned by students in the validation sample are reported in Table 15. These indicate that the mathematics achievement level of the group was average. The mean

values reported were 663.4, 678.4, 659.0, and 666.0, respectively, for the Concepts of Number, Mathematics Computation, Mathematics Applications and Total Mathematics tests. The mean scaled scores earned by the national sample of sixth graders who participated in the spring 1982 standardization and norming of the Stanford were 656, 674, 653, and 660, respectively.

Comparison of Estimated and Actual Scaled Scores. The four mathematics scaled scores estimated for students in the validation sample were compared in several ways to those actually earned by these students. First, a Pearson product-moment correlation was computed. The values reported in Table 16 are inflated estimates of the relationship between estimated and actual scaled scores because of the part-whole relationship between them. However, these values, ranging between .937 and .976, suggest a high degree of relationship between pairs.

The accuracy of pairs of scaled scores was further evaluated by computing a root mean square difference for each pair. The values obtained are also reported in Table 16. A root mean square difference is an index reflecting the overall differences between pairs of scaled scores. To facilitate comparisons across tests, the values have been scaled by their variances. The values reported for the pairs of scaled scores under study here were .362, .302, .337, and .221, respectively, for Concepts of Number, Mathematics Computation, Mathematics Applications, and Total Mathematics. These indicate that, overall, the estimated scaled scored were quite similar to those actually earned by the students in the validation sample. A smaller scaled root mean square difference was observed for the Total Mathematics test. This result is consistent with the fact that this test contains more items than any of the subtests.

Finally, the decrease in the accuracy of the estimated scaled scores was evaluated by examining the standard error of measurement associated with each scaled score pair. For the total mathematics test, the standard error of measurement decreased only three scaled-score points in the mid-range of the scaled score-distribution. This value was five scaled-score points for the three subtests. Such small decreases in the standard error of measurement suggest that only a small amount of accuracy is lost when using a small set of carefully selected items to estimate student performance at the subtest level.

IMPLICATIONS FOR CALIFORNIA'S COMPREHENSIVE ASSESSMENT SYSTEM

The results of this study have a variety of implications for California's Comprehensive Assessment System. First, the study lends support to one of the basic premises of the system, namely the assumption that some of the testing programs currently in place in California evaluate common skills. This assumption is essential to the concept of a consolidated and unified assessment system for California.

Second, the study points out the need for well-defined and articulated content domains that include all the important educational outcomes in the state. Developing such a structure will take time and the cooperative efforts of the educational community, the state, and test publishing companies like The Psychological Corporation.

Third, the study demonstrates the feasibility of using multi-purpose testing as a methodology for accomplishing needed reforms in the state's current testing and assessment system. As a demonstration of this methodology, a set of multi-purpose items was identified and then used to generate accurate norm-referenced test information.

References

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R.D., & Mislevy, R.J. (1981). An item response model for matrix-sampling data: The California grade-three assessment. In D. Carlson (Ed.), New Directions for Testing and Measurement. San Francisco, CA: Josey-Bass.

Bock, R.D., Mislevy, R.J., & Woodson, C.E. (1982). The next stage in educational assessment. Educational Researcher, 11, 4-11.

California State Department of Education (1980). Survey of Basic Skills: Grade 3, rationale and content. Sacramento, CA: California State Department of Education.

California State Department of Education (1982a). Survey of Basic Skills: Grade 6, rationale and content. Sacramento, CA: California State Department of Education.

California State Department of Education (1982b). Student achievement in California schools: 1981-82 annual report. Sacramento, CA: California State Department of Education.

California State Department of Education (1984a). A technical description of the procedures used in calculating school-level scaled scores for the Survey of Basic Skills. Sacramento, CA: California State Department of Education.

California State Department of Education (1984b). Survey of Academic Skills: Grade 8, rationale and content (Mathematics). Sacramento, CA: California State Department of Education.

California State Department of Education (1984c). Survey of Academic Skills: Grade 8, rationale and content (Reading and Written Language). Sacramento, CA: California State Department of Education.

Dempster, A.P., Laird, N.M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Gardner, E.F., Rudman, H.C., Karlsen, B., & Merwin, J. (1982). Stanford Achievement Test, Seventh Edition. Cleveland, OH: The Psychological Corporation.

Gardner, E.F., Rudman, H.C., Karlsen, B., & Merwin, J. (1983). Stanford Achievement Test, Seventh Edition, Index of Instructional Objectives. Cleveland, OH: The Psychological Corporation.

Jolly, S.J., & Gramenz, G.W. (1984). Customizing a norm-referenced achievement test to achieve curricular valididty: A case study. Educational Measurement, 3(3), 16-18.

Madaus, G.F., Kellaghan, T., Rakow, E.A., & King, D.J. (1979). The sensitivity of measures of school effectiveness. Harvard Educational Review, 49(2), 207-230.

Mehrens, W.A. (1984). National tests and local curriculum: Match or mismatch? Educational Measurement, 3(3), 9-15.

Mislevy, R.J. (1983). Item response models for grouped data. Journal of Educational Statistics, 8, 271-288.

Mislevy, R.J., & Bock, R.D. (1984). BILOG: Item analysis and test scoring with binary logitic models. Mooresville, IN: Scientific Software, Inc.

Popham, W.J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogishe Institute, 1960.